Smart Rating for Electronics Gadgets Data

Navya Bharathi Alla

Advisor : Prof. Jeongkyu Lee

Data Mining

University of Bridgeport, Bridgeport, CT

Abstract

•Aggregator websites of Electronic gadgets listings decide if they want to pick a particular listing from an e-commerce website and display on their own website. Selection of the right items to list is essential to customers selecting products for purchase, therefore leading to higher revenues.

•Many e-commerce websites do not collect user ratings, or do not provide this data to our website. Hence, it is difficult to determine whether to display the product or not. Simply removing all items without ratings could lead to high opportunity costs, hence, another solution is required

•Predict the *High - Low* rating for any new electronic gadget listing to be introduced on the aggregator website monitored by bargain.in. The same model can also be used to predict the High - Low Rating on websites that do not support the feature of Average Rating.

Introduction

Business Objective: Increase coverage of listings that can be included in our aggregator website, by including those items that do not have ratings. This has the potential to improve our sales revenues by selection of appropriate items that are likely to be purchased by customers, and also avoids the opportunity cost of simply not listing such products.

Data Mining Objective: Predict the *High - Low* Rating for any new electronic gadget listing to be introduced on the aggregator website monitored by our website. The same model can also be used to predict the High - Low Rating on websites that do not support the feature of Average Rating.

Dataset and Data Preparation

• Data is being collected online from the website with details about the electronic gadgets listed on the ecommerce websites.

•The data required the following transformations in order to perform the modeling.

Column	Transformation
Brand	Created categorical variables for each brand. Reduced categories by studying the pivot table of avRating (output) with brand categories.
AvRating	Created Binned variables: Values less than 2 -> LOW otherwise HIGH
ShippingPeriod	Missing values were generated used KNN – prediction from available datapoints using sitenames, category, instock and freeshipping as inputs (see Appendix B)
SiteName	Created categorical variables for each website.
Category	Created categorical variables for each category.

AnalyticalMethod

Using Logistic regression method a model is created to predict high-low rating of the test data where high rating is considered as success measure and cut off point at 0.6 is kept is used to predict the probability of success. 10.25% prediction error observed in Naïve method is improved to prediction error of 2.91% in Logistic regression method.

Validation Data scoring - Summary Report (Logistic regression)

Cut off Prob.Val. for Success (Updatable) 0.6

Classification Confusion Matrix

Actual Class \ Predicted Class	High	Low
High	1484	7
Low	42	148

Error Report

Class	# Cases	# Errors	% Error
High	1491	7	0.47
Low	190	42	22.11
Overall	1681	49	2.91

Naïve Approach

Let's first assess the benchmark Naïve model. This is a fairly straightforward model wherein the prediction is based on majority and the error rates are around 10.25%

Summary report (Naïve)

Actual Class \ Predicted class	High	Low	Grand Total
High	3771	0	3771
Low	431	0	431
Grand Total	4202	0	4202
% Error		10.2570205	

Conclusion

- **The model is of utmost use to product owners who intend to launch their product via these websites. They can assess on which of the sites will their products garner a high rating and which of the sites will underrate their products.**
- **The relative advantages of each of the methodologies came to the fore when we started building the models. In our assessment, we recommend that we use Naïve bayes in case of a small training set, while with bigger sets, we can go for KNN of logistic regression.**
- **There can be some overfitting in the logistic regression model, so running multiple models for the same data set is always useful**

References

- http://www.ats.ucla.edu/stat/sas/topics/logistic.htm
- http://www.ats.ucla.edu/stat/r/dae/logit.htm
- http://en.wikipedia.org/wiki/Logistic_regression
- https://archive.ics.uci.edu/ml/datasets.html

Acknowledgement

Our sincere thanks to Prof. Jeongkyu Lee for assigning us this real world challenging project and for his valuable inputs and guidance